Attorney Docket No.: 0160112

UNITED STATES PATENT APPLICATION

FOR

SIGNAL DECOMPOSITION OF VOICED SPEECH FOR CELP SPEECH CODING

INVENTOR:

YANG GAO

CERTIFICATE OF EXPRESS MAILING

I hereby certify that this correspondence is being deposited with the United States Postal Service "Express Mail Post Office to addressee" Service under 37 C.F.R. Sec. 1.10 addressed to: Commissioner for Patents, P. O. Box 1450, Alexandria, VA 22313-1450, on 3/11/04

Express Mailing Label No.: EV420420970US

Lori Lapidario   Lori Lapidario
Name   Signature

PREPARED BY:

FARJAMI & FARJAMI LLP
26522 La Alameda Ave., Suite 360
Mission Viejo, California 92691

(949) 282-1000
Customer No. 25700

25700
PATENT TRADEMARK OFFICE

SIGNAL DECOMPOSITION OF VOICED SPEECH FOR CELP SPEECH CODING

RELATED APPLICATIONS

The present application claims the benefit of United States provisional application serial number 60/455,435, filed March 15, 2003, which is hereby fully incorporated by reference in the present application.

The following co-pending and commonly assigned U.S. patent applications have been filed on the same day as this application, and are incorporated by reference in their entirety:

United States Patent Application Serial Number ___________, "VOICING INDEX CONTROLS FOR CELP SPEECH CODING," Attorney Docket Number: 0160113.

United States Patent Application Serial Number ___________, "SIMPLE NOISE SUPPRESSION MODEL," Attorney Docket Number: 0160114.

United States Patent Application Serial Number ___________, "ADAPTIVE CORRELATION WINDOW FOR OPEN-LOOP PITCH," Attorney Docket Number: 0160115.

United States Patent Application Serial Number ___________, "RECOVERING AN ERASED VOICE FRAME WITH TIME WARPING," Attorney Docket Number: 0160116.

BACKGROUND OF THE INVENTION

1. FIELD OF THE INVENTION

The present invention relates generally to speech coding and, more particularly, to Code Excited Linear Prediction (CELP) for wideband speech coding.

2. RELATED ART

Generally, a speech signal can be band-limited to about 10 kHz without affecting its perception. However, in telecommunications, the speech signal bandwidth is usually limited much more severely. It is known that the telephone network limits the bandwidth of the speech signal to between 300 Hz to 3400 Hz, which is known as the "narrowband". Such band-limitation results in the characteristic sound of telephone speech. Both the lower limit at 300Hz and the upper limit at 3400 Hz affect the speech quality.

In most digital speech coders, the speech signal is sampled at 8 kHz, resulting in a maximum signal bandwidth of 4 kHz. In practice, however, the signal is usually band-limited to about 3600 Hz at the high-end. At the low-end, the cut-off frequency is usually between 50 Hz and 200 Hz. The narrowband speech signal, which requires a sampling frequency of 8 kb/s, provides a speech quality referred to as toll quality. Although this toll quality is sufficient for telephone communications, for emerging applications such as teleconferencing, multimedia services and high-definition television, an improved quality is necessary.

The communications quality can be improved for such applications by increasing the bandwidth. For example, by increasing the sampling frequency to 16 kHz, a wider bandwidth, ranging from 50 Hz to about 7000 Hz can be accommodated, which is referred to as the "wideband". Extending the lower frequency range to 50 Hz increases naturalness, presence and comfort. At the other end of the spectrum, extending the higher frequency range to 7000 Hz increases intelligibility and makes it easier to differentiate between fricative sounds.

Digitally, speech is synthesized by a well-known approach known as Analysis-By-Synthesis (ABS). Analysis-By-Synthesis is also referred to as closed-loop approach or waveform-matching approach. It offers relatively better speech coding quality than other approaches for medium to high bit rates. A known ABS approach is the so-called Code Excited Linear Prediction (CELP). In CELP coding, speech is synthesized by using encoded excitation information to excite a linear predictive coding (LPC) filter. The output of the LPC filter is compared against the voiced speech and used to adjust the filter parameters in a closed loop sense until the best parameters based upon the least error is found. The problem with this approach is that the waveform is difficult to match in the presence of noise in the speech signal.

Another method of speech coding is the so-called harmonic coding approach. Harmonic coding assumes that voiced speech is approximated by a series of harmonics. And when all the harmonics are added together, a quasi-periodic waveform appears. Thus working on the principle that voiced speech is quasi-periodic, it is easier to match voiced speech using prior art Harmonic coding approaches.

Waveform matching or harmonic coding is easier for periodic speech components than non-periodic speech components. This is because non-periodic speech signal is random-like and broadband thus would not fit in the basic harmonic model. However, the harmonics approximation approach may be too simplistic for real voiced signals because real voiced signals include irregular (i.e. noise) components. Thus, high quality waveform-matching becomes difficult even for voiced speech, because of significant irregular components that may exist in the voiced signal especially for wideband speech signal. These irregular components usually occur in the high frequency areas of the wideband voice signals but, may also be present throughout the voice band.

The present invention addresses the above voiced speech issue because real world speech signal may not be periodic enough so that a perfect waveform matching becomes difficult. .

SUMMARY OF THE INVENTION

In accordance with the purpose of the present invention as broadly described herein, there is provided systems and methods for improving quality of synthesized speech by decomposing input speech into a voiced portion and a noise portion. The voice portion is coded using CELP methods thus allocating most of the bit budget to the voiced speech for true quality reproduction. This portion (voiced) covers mostly the low to mid frequency range. The noise portion of the input speech is allocated the least bit budget and may be estimated at the decoder since it contains minimal voiced speech components. The noise portion is usually in the high frequency range.

The decomposition of the input speech into the two portions is frequency dependent and is adaptive to the input speech. In one embodiment, the separation occurs after background noise has been removed from the input speech. The decomposition may be accomplished using a lowpass/highpass filter combination. The information regarding bandwidth of the lowpass/highpass may be presented to the decoder to facilitate reproduction of the noise portion of the speech. The information about the appropriate filter cut-off frequency may be provided to the decoder in the form of voicing index, for example.

The decoder may synthesize the input speech by using a CELP process on the voiced portion and injecting noise to represent the noise portion.

These and other aspects of the present invention will become apparent with further reference to the drawings and specification, which follow. It is intended that all such additional systems, methods, features and advantages be included within this description, be within the scope of the present invention, and be protected by the accompanying claims.

BRIEF DESCRIPTION OF DRAWINGS

Figure 1 is an illustration of the frequency domain characteristics of a voiced speech signal.

Figure 2 is an illustration of separation of speech residual (or excitation) into a voiced component and a noise component in accordance with an embodiment of the present invention.

Figure 3 is an illustration of synthesis of voiced speech from voiced components in accordance with an embodiment of the present invention.

DETAILED DESCRIPTION

The present application may be described herein in terms of functional block components and various processing steps. It should be appreciated that such functional blocks may be realized by any number of hardware components and/or software components configured to perform the specified functions. For example, the present application may employ various integrated circuit components, e.g., memory elements, digital signal processing elements, transmitters, receivers, tone detectors, tone generators, logic elements, and the like, which may carry out a variety of functions under the control of one or more microprocessors or other control devices. Further, it should be noted that the present application may employ any number of conventional techniques for data transmission, signaling, signal processing and conditioning, tone generation and detection and the like. Such general techniques that may be known to those skilled in the art are not described in detail herein.

Figure 1 is an illustration of the frequency domain characteristics of a voiced speech signal. In this illustration, the spectrum domain in the wideband extends from slightly above 0 Hz to around 7.0 kHz. Although the highest possible frequency in the spectrum ends at 8.0 kHz (i.e. Nyquist folding frequency) for a speech signal sampled at 16 kHz, this illustration shows that the energy is almost zero in the area between 7.0 kHz to 8.0 kHz. It should be apparent to those of skill in the arts that the ranges of signals used herein are for illustration purposes only and that the principles expressed herein are applicable to other signal bands.

As illustrated in Figure 1, the speech signal is quite harmonic at lower frequencies, but at higher frequencies the speech signal does not remain as harmonic because the probability of having noisy speech signal increases as the frequency increases. For instance, in this illustration the speech signal exhibits traits of becoming noisy at the higher frequencies, e.g., above 5.0 kHz. If we call this frequency point (5.0 kHz) the voicing cut-off frequency, this voicing cut-off frequency could vary from 1 kHz until 8 kHz for different voiced signal. The noisy signal makes waveform matching at higher frequencies very difficult. Thus, techniques like ABS coding (e.g. CELP) becomes unreliable if high quality speech is desired. For example, in a CELP coder, the synthesizer is designed to match the original speech signal by minimizing the error between the original speech and the synthesized speech. A noisy signal is unpredictable thus making error minimization very difficult.

Given the above problem, the present invention decomposes the speech signal into two portions, namely a voiced (or major) portion and a noisy portion. The voiced portion comprises the region from low to high frequency (e.g., 0-5 kHz in Figure 1) where the speech signal is relatively harmonic thus amenable to analysis-by-synthesis methods. Note that noise may be present in the voiced portion however, speech predominates in this region for voiced speech.

The noise portion may comprise random speech signal. Since most noise-like components are predominant in the high frequency region (as shown in Figure 1), in one embodiment, the signal decomposition could be done by adaptive low-pass filtering and/or high-pass filtering of the speech residual signal.

Figure 2 is an illustration of separation of speech residual (or excitation) into a voiced component and a noise component in accordance with an embodiment of the present invention. In this illustration, Input Speech 201 is processed through LPC analysis 204 and Inverse filter 202 to generate Residual 205. Residual 205 is subsequently processed through an appropriate Lowpass filter 206 to generate Voiced Residual 207. Lowpass 206 may be adaptively selected from a group of preprogrammed low-pass filters that is known to both the encoder (e.g. 200) and the decoder (e.g. 300). For instance, the filter structure may be fixed but the bandwidth may vary depending on several factors, which may be determined through Voicing Analysis 208, such as: Pitch correlation, gender of the speaker, etc. Thus, the speech signal decomposition of the present invention is adaptive to speech.

In an embodiment, normalized Pitch correlation may be used to select an appropriate filter bandwidth. In such a case, the logic may be such that when normalized pitch correlation is close to 1 (one), the filter bandwidth is almost at infinity. This is because in such a case (i.e. pitch correlation close to one), the waveform of Input Speech 201 more closely resembles a harmonic model throughout the frequency band of interest. On the other extreme, the bandwidth selected may approach zero as pitch correlation approaches zero. In this case, i.e. pitch correlation close to zero, the waveform of Input Speech 201 more closely resembles an unvoiced speech model thus more characteristically resembles noise. Thus, the task is to find an appropriate relationship between normalized pitch correlation and filter bandwidth.

The selected filter may be communicated to the Decoder 300 using a group of bits that when decoded at the decoder indicates which filter was selected at the encoder. This group of bits may be referred to as the voicing index.

In accordance with one embodiment, a voicing index defines a plurality of low pass filters, such as seven or eight different low pass filters, for which three (3) bits are transmitted from the encoder to the decoder. In like manner, four (4) bits may be used when there are between eight and sixteen filter selections available. Of course, the number of different filters and the method of communicating the selected filter parameters depends on the complexity and accuracy of the implementation.

In one embodiment, the voiced portion 207 of the speech signal is encoded using CELP process in block 210. CELP processing may be desirable over Harmonic coding because it should provide better quality speech with higher bit budget. Harmonic coding is generally good for low frequency applications because the requirement for aggregate rate (bit budget) is less than for the CELP model. However, it is generally difficult for Harmonic models to reproduce very high quality speech in the presence of some noise since it may not be possible to completely separate noise from the voiced speech. Moreover, increasing the bit budget to relatively high bit-rate for a harmonic model does not improve the quality of the reproduction as much as a CELP model.

On the other hand, the CELP coder may still generate high quality speech even in the presence of some noise by simply increasing the bit budget. Thus, a CELP or similar high quality coder is preferably used on the voiced portion to improve the quality of the synthesized speech.

In one embodiment, CELP coder 210 spends the available bits to code the voiced residual portion 207 at the encoder and transmits the coded information, such as LPC parameters, pitch, energy, excitation, etc. to the decoder 300. At the decoder 300, the coded information is decoded and used to synthesize the voiced portion 309 (See Figure 3), and the noisy portion is estimated using random noise excitation.

The noise portion, because it is hard to waveform match, does not have to be coded. Moreover, the noise portion may be represented by an excitation and an LPC filter envelope because once the LPC envelope is removed, the excitation is characteristically flat. Thus, the noise portion need not be coded because it could easily be estimated with knowledge of the LPC filter parameters and the magnitude of the voiced speech portion at the cutoff frequency of the lowpass filter 201.

The selected filter parameters may be communicated to the Decoder 300 using a group of bits (e.g. the voicing index) that when decoded at the decoder indicates which filter was selected for the noise portion. For example, if there are up to eight different filters available, then three bits may be used to indicate the selected filter. In like manner, four bits may be used when there are between eight and sixteen filter selections available. Of course, the number of different filters and the method of communicating the selected filter parameters depends on the complexity and accuracy of the implementation.

In one embodiment, the noise portion is not coded because an excitation (e.g. white noise) may be passed through the selected high-pass filter and LPC synthesis filter at the decoder 300 to synthesize the noise portion, which may then be added to the synthesized voiced portion to form Output Speech 301. The noise portion needs to be normalized to the magnitude of the voiced portion at the cutoff frequency of the lowpass filter at the decoder.

Other embodiments of the invention may use other convenient method to separate the voiced portion from the noise portion. For instance, a harmonic model may be used. In the harmonic model, the true input speech may be compared to the harmonic prediction of the speech and the model that gives the least error (e.g. Mean Square Error) may be selected to represent the voiced portion.

In one or more embodiments, each low pass filter implemented for separation of the voiced portion from the noise portion, there is a corresponding high pass filter. At the decoder side, the voicing index value indicates which low pass filter (thus its corresponding high pass filter) was used in separating the voiced portion from the noisy portion and this knowledge is used to synthesize the input speech signal. Figure 3 is an illustration of synthesis of speech at the decoder in accordance with an embodiment of the present invention.

In this illustration, the voiced portion is decoded at block 304 based on CELP parameters received from the encoder. The generated signal is adaptively filtered in block 308, using the adaptive lowpass filter parameters obtained from the voicing index, to generate the voiced portion 309. Further, a noise generator 302 may be utilized at the decoder to generate random noise, which is then processed through the high pass filter 306. Highpass filter 306 is also adaptive and is based on information obtained from the voicing index and is the corresponding one of lowpass filter 308.

In block 310, the signal energy of the noise portion is adjusted proportionately with the generated voiced potion, so that the energy remains flat when the voiced component and the noise component are summed in block 312. In one embodiment, the noise portion 311 may be generated using a highpass filter, e.g. 306, which may be implemented with the transfer function (1-Lowpass 308). Thus, after selection of an appropriate filter bandwidth, Voiced portion 309 and Noise portion 311 may be readily generated using lowpass and highpass filters, respectively.

After summation, in block 312, of voiced portion 309 and noise portion 311, the resulting speech signal is processed through synthesis filter 314 and post processing block

316 to obtain the output speech signal, 301, which is the synthesized speech.

Although the above embodiments of the present application are described with reference to wideband speech signals, the present invention is equally applicable to narrowband speech signals.

The methods and systems presented above may reside in software, hardware, or firmware on the device, which can be implemented on a microprocessor, digital signal processor, application specific IC, or field programmable gate array ("FPGA"), or any combination thereof, without departing from the spirit of the invention. Furthermore, the present invention may be embodied in other specific forms without departing from its spirit or essential characteristics. The described embodiments are to be considered in all respects only as illustrative and not restrictive.